Abstract of the Disclosure

A method and system for identifying groups of pages of common interest from a collection of hyper-linked pages are disclosed. A plurality of community cores are identified from the collection where each core includes first and second sets of pages, and each page in the first set points to every page in the second set. Each identified core is expanded into a full community which is a subset of the pages regarding a particular topic. The identification community cores is based on the analysis of the Web graph in which the communities correspond to instances of Web subgraphs. Extraneous pages are then pruned to improve the quality of the resulting communities.